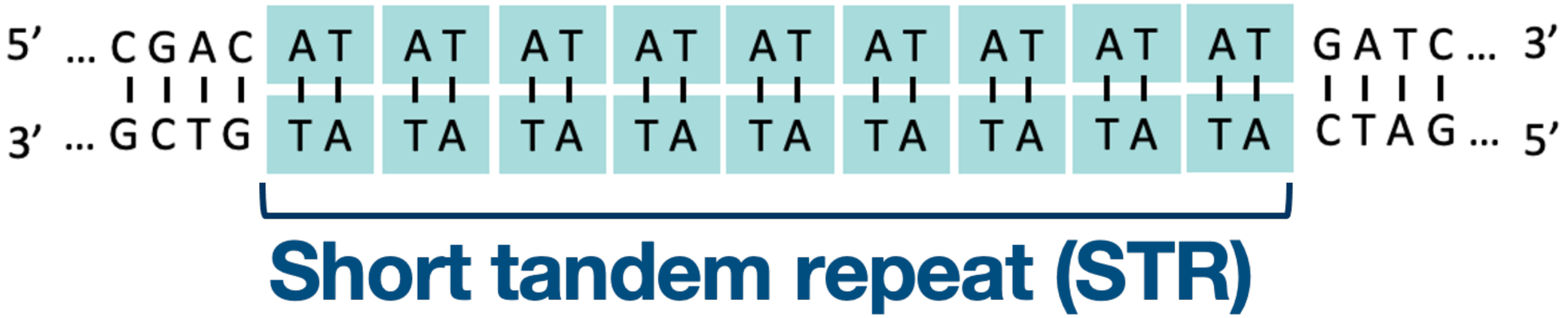
Analysing short tandem repeats in genomic sequences

What, why, how?

BIO392

Short tandem repeats (a.k.a. microsatellites)

Repetitions of 1-6 nt DNA motifs



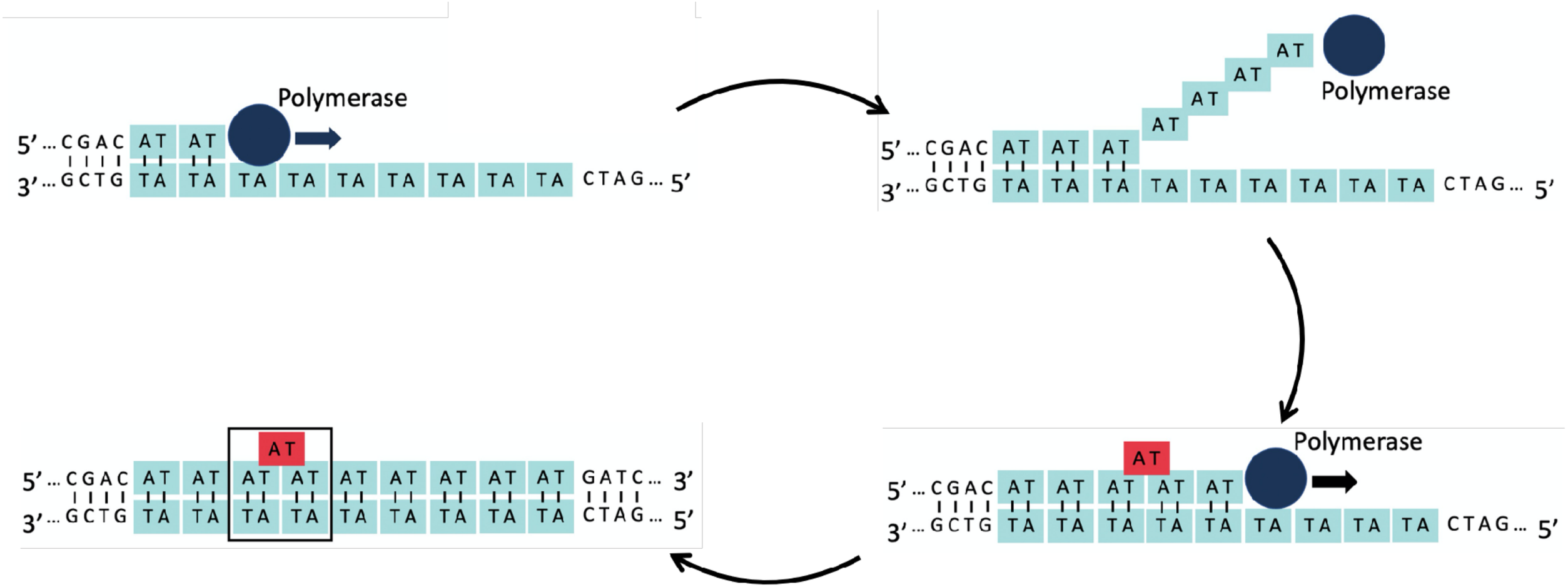
Motif: the repeating unit (‘AT’ in this case)

Motif length: length of the repeating unit

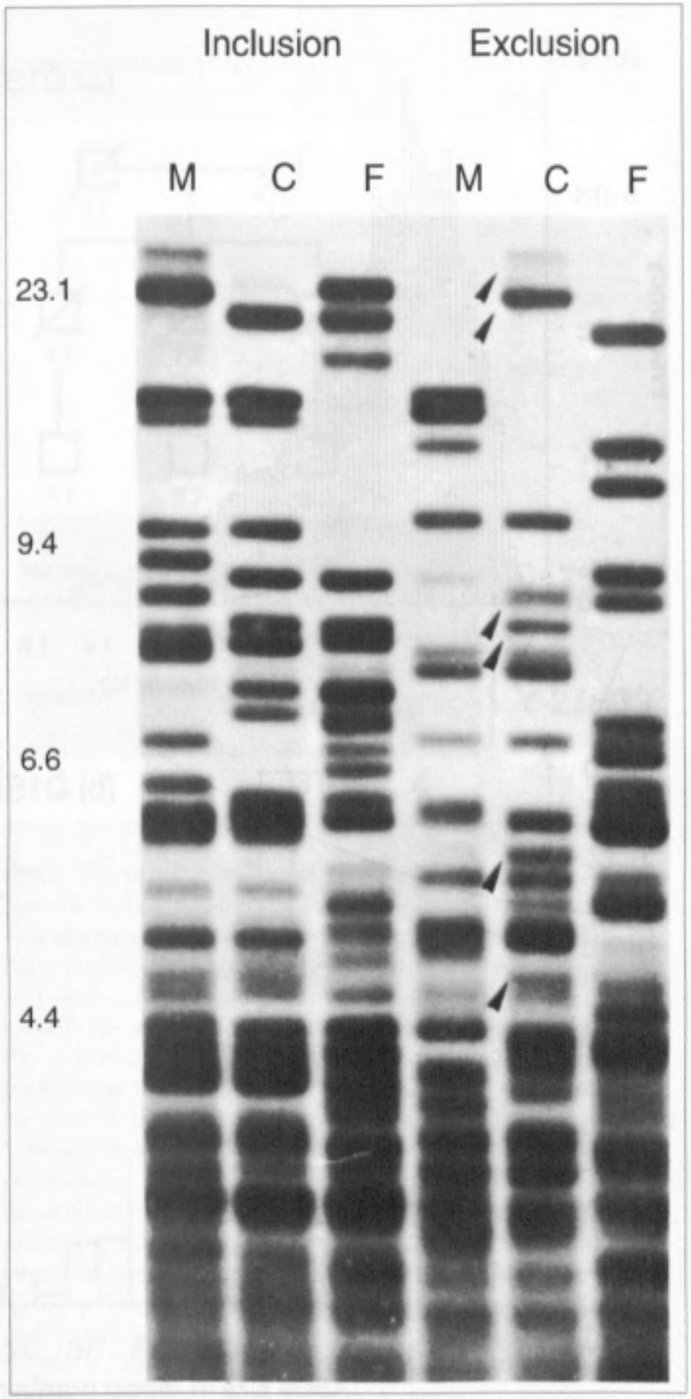
Copy number: number of times the unit is repeated

STRs have mutation rates up to 10’000 times higher than point mutations

Slipped-strand mispairing

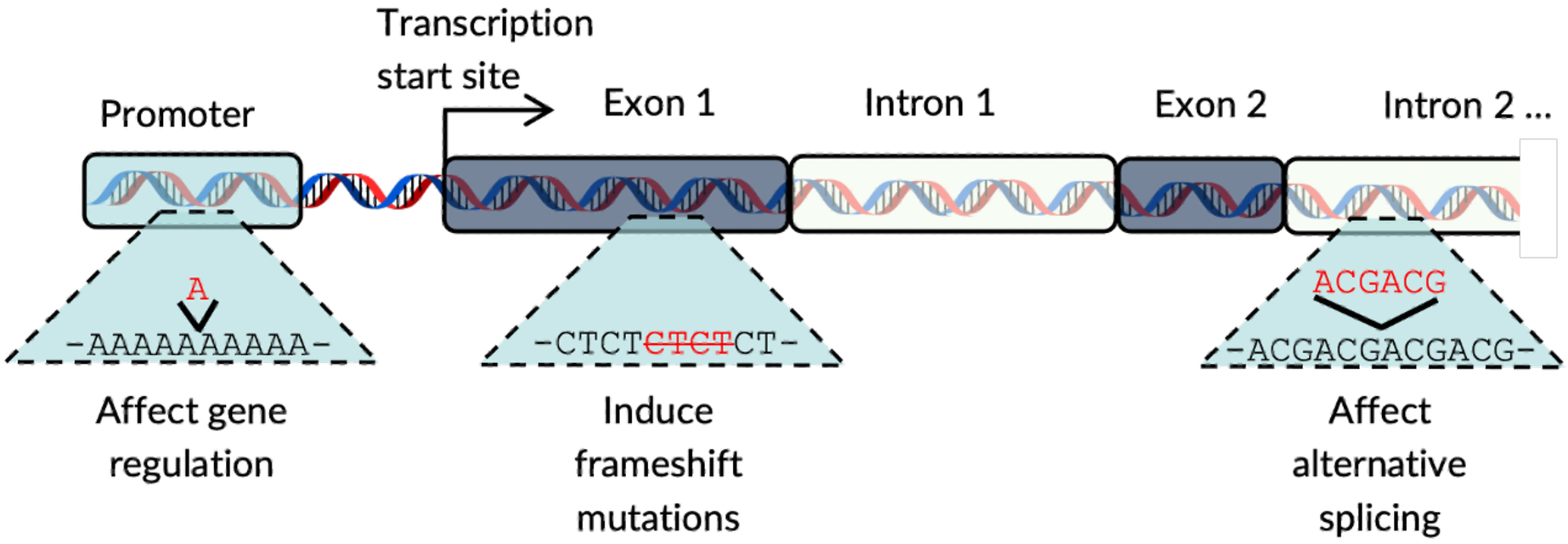


Classic use case: paternity testing

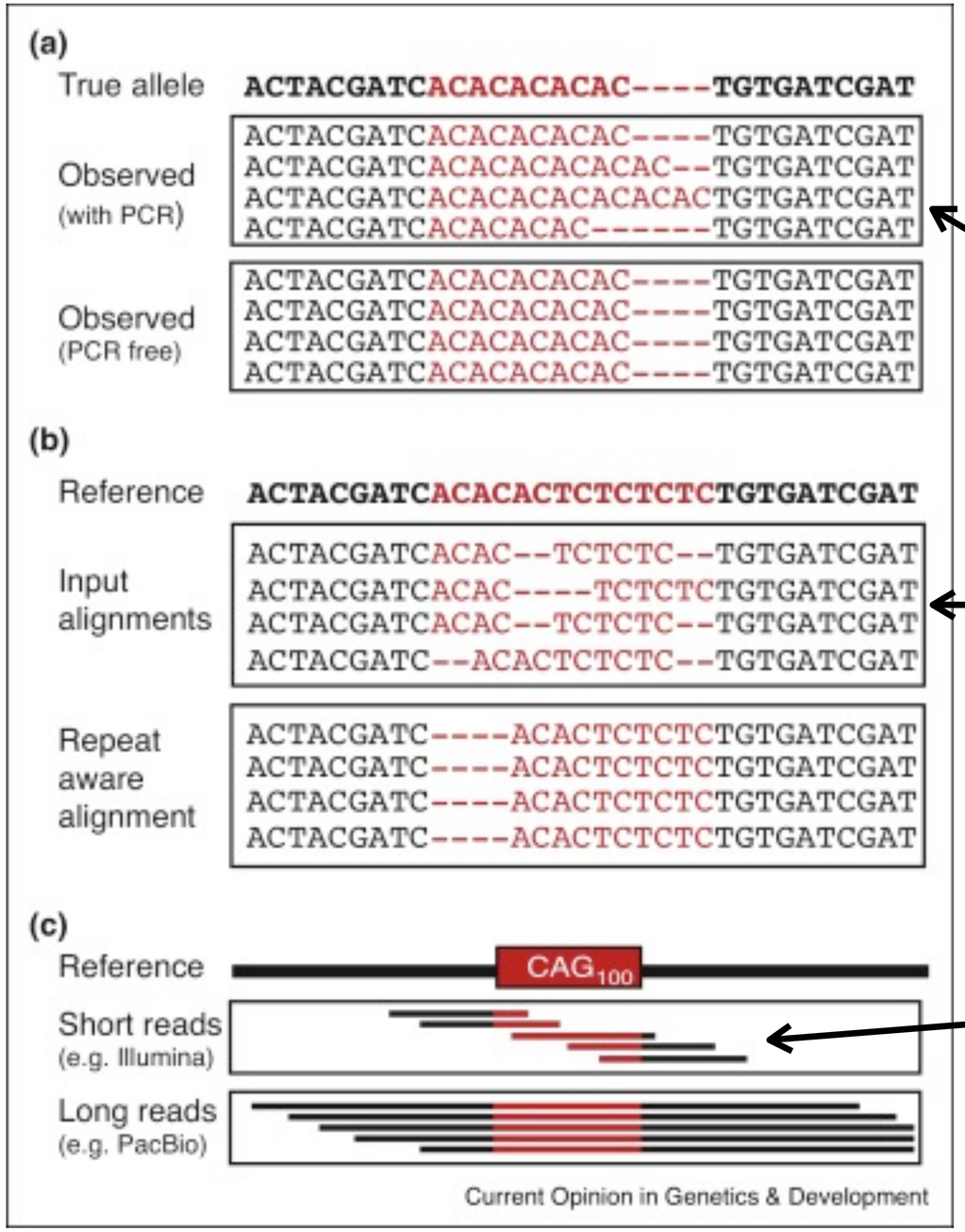


Pena, S. D. J., & Chakraborty, R. (1994). Paternity testing in the DNA era. Trends in Genetics, 10(6), 204–209. doi:10.1016/0168-9525(94)90257

Functional consequences of STR mutations



Why STRs are challenging to genotype



PCR induces mistakes during sequencing (addressed by PCR-free methods)

Not all alignment algorithms are ‘repeat aware’

Short reads do not always span longer STRs, which makes determining their length very hard

Gymrek, Melissa. ‘A Genomic View of Short Tandem Repeats’. *Current Opinion in Genetics and Development* 44 (1 June 2017): 9–16. https://doi.org/10.1016/j.gde.2017.01.012.

We need special tools to genotype STRs!

e.g. Table 1 in https://doi.org/10.1016/j.gde.2017.01.012

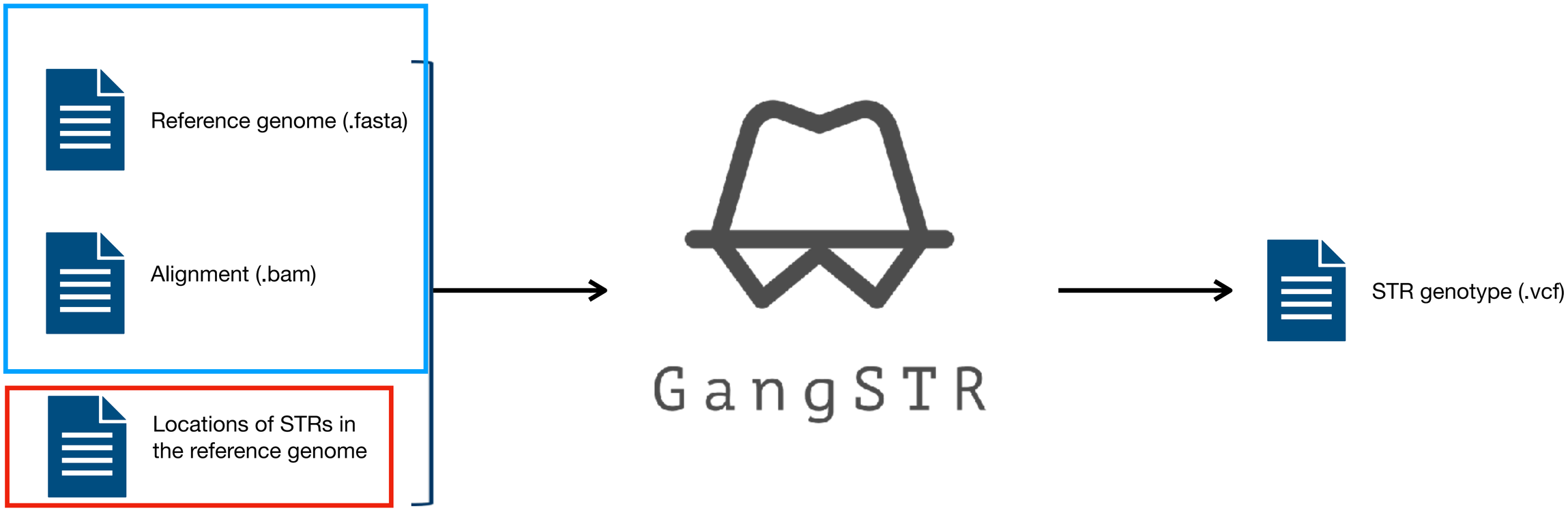
- We will use GangSTR today: https://github.com/gymreklab/GangSTR

GangSTR

Mousavi, Nima, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. 'Profiling the Genome-Wide Landscape of Tandem Repeat Expansions'. *Nucleic Acids Research* 47, no. 15 (5 September 2019): e90–e90. https://doi.org/10.1093/NAR/GKZ501.

GangSTR intput/output

These files we have!
Reference: available online
Alignment: output of sequencing experiment



This file we need to generate ourselves!

Detecting (short) tandem repeats in genomic sequence

- Computationally expensive for large sequences (especially imperfect repeats)
- Many different algorithms exists, each with different strengths/weaknesses

Anisimova, Maria, Julija Pečerska, and Elke Schaper. ‘Statistical Approaches to Detecting and Analyzing Tandem Repeats in Genomic Sequences’. *Frontiers in Bioengineering and Biotechnology* 3, no. MAR (2015): 31. https://doi.org/10.3389/FBIOE.2015.00031.

Tandem repeat annotation library (TRAL)

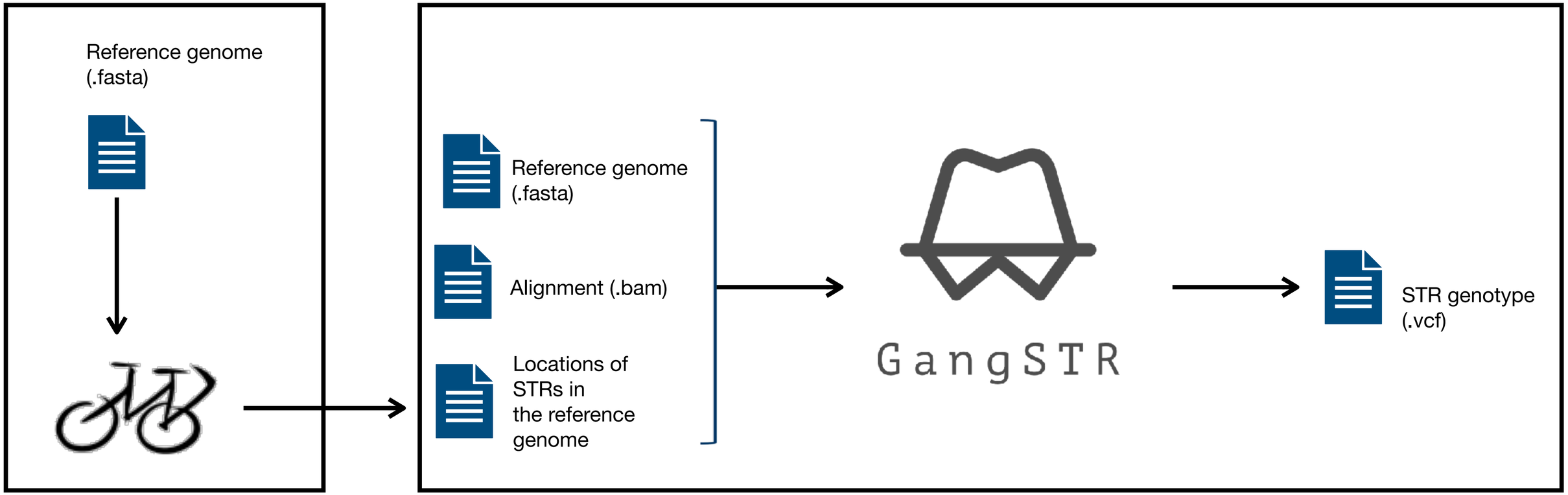
- A python library aimed at addressing challenges in repeat detection:
  - Allows for the running of several repeat detection algorithms on the input sequence
  - Calculates scores for each detected repeat for filtering
  - Can make the set of detected repeats non-redundant

Schaper, Elke, Alexander Korsunsky, Jūlija Pečerska, Antonio Messina, Riccardo Murri, Heinz Stockinger, Stefan Zoller, Ioannis Xenarios, and Maria Anisimova. ‘TRAL: Tandem Repeat Annotation Library’. *Bioinformatics* 31, no. 18 (15 September 2015): 3051–53. https://doi.org/10.1093/bioinformatics/btv306.

Summary

- Short tandem repeats can mutate rapidly, which can have functional consequences
- Repeat *genotyping*: determining how long repeats are in a given sequencing sample (e.g. using GangSTR)
- Repeat *detection*: determining where repeats are located and how long they are in the reference sequence (e.g. using TRAL)

Workflow for today



Task 1

STR background reading

Gymrek, Melissa. ‘A Genomic View of Short Tandem Repeats’. *Current Opinion in Genetics and Development* 44 (1 June 2017): 9–16. https://doi.org/10.1016/j.gde.2017.01.012.

- Read the following sections of ‘A genomic view of short tandem repeats’:
  - Abstract + Introduction
  - Genotyping STRs from high-throughput sequencing data

- Afterwards, you should be able to answer the following questions:
  - Why is STR variation relevant to health and disease?
  - What are some of the challenges in analysing STRs from NGS data?

Task 2

Introduction to TRAL

Schaper, Elke, Alexander Korsunsky, Jūlija Pečerska, Antonio Messina, Riccardo Murri, Heinz Stockinger, Stefan Zoller, Ioannis Xenarios, and Maria Anisimova. ‘TRAL: Tandem Repeat Annotation Library’. *Bioinformatics* 31, no. 18 (15 September 2015): 3051–53. https://doi.org/10.1093/bioinformatics/btv306.

- Read ‘TRAL: tandem repeat annotation library’

- Afterwards, you should be able to answer the following questions:
  - Why should you use multiple tandem repeat detection algorithms to look for repeats in biological sequence?
  - What different functionalities does TRAL provide?

Task 3

STR detection in the *APC* gene

- Set up the ‘BIO392_STRs’ conda environment and activate it (see appendix 1)
- Start a jupyter server and open ‘scripts/BIO392_TRAL.ipynb’ (see appendix 2)
- Follow along with the steps in BIO392_TRAL.ipynb

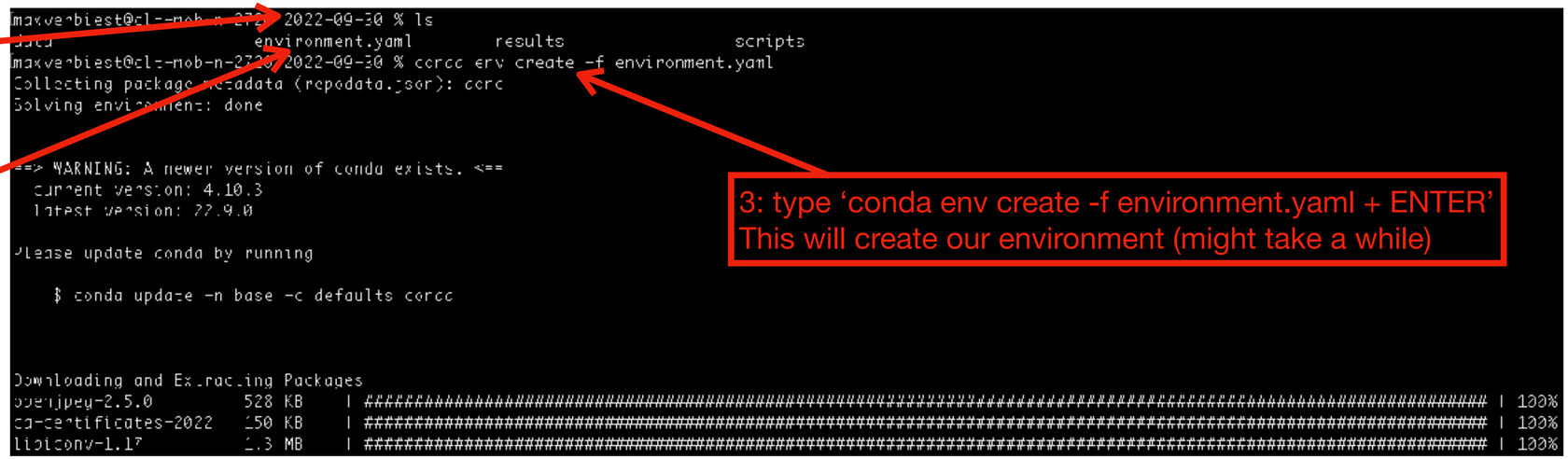
Appendix 1

Setting up the BIO392_STRs conda env

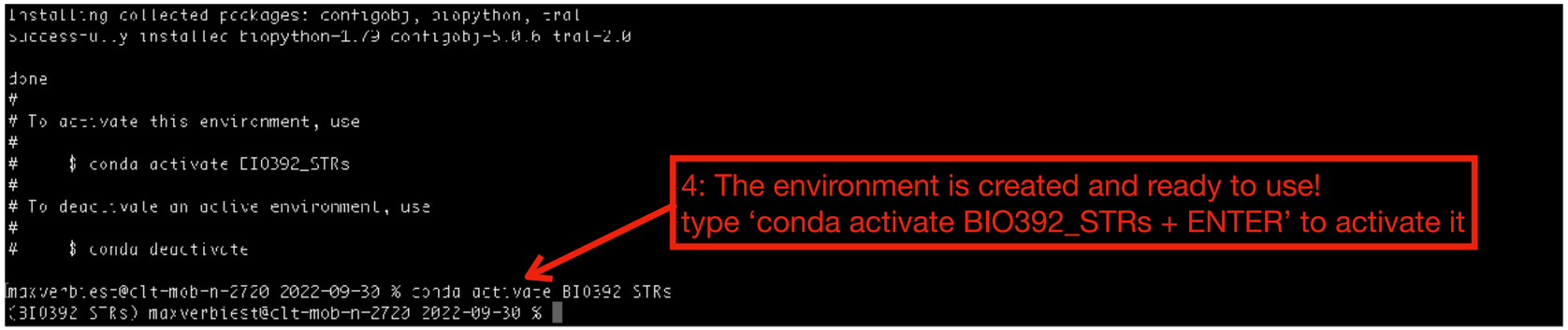
1: Make sure you are in the right directory

2: type ‘ls + ENTER’ you should see this file

3: type ‘conda env create -f environment.yaml + ENTER’ This will create our environment (might take a while)



...



4: The environment is created and ready to use! type ‘conda activate BIO392_STRs + ENTER’ to activate it

Appendix 2

Starting a jupyter server